An Information Theoretic Evaluation Metric for Strong Unlearning

Dongjae Jeon[1*], **Wonje Jeung**[1*], **Taeheon Kim**[2], **Albert No**[1†], **Jonghyun Choi**[2†]

[1]Yonsei University
[2]Seoul National University
{dongjae0324, specific0924, albertno}@yonsei.ac.kr, {thkim0305, jonghyunchoi}@snu.ac.kr

Abstract

Machine unlearning (MU) aims to remove the influence of specific data from trained models, addressing privacy concerns and ensuring compliance with regulations such as the "right to be forgotten." Evaluating strong unlearning, where the unlearned model is indistinguishable from one retrained without the forgetting data, remains a significant challenge in deep neural networks (DNNs). Common black-box metrics, such as variants of membership inference attacks and accuracy comparisons, primarily assess model outputs but often fail to capture residual information in intermediate layers. To bridge this gap, we introduce the Information Difference Index (IDI), a novel white-box metric inspired by information theory. IDI quantifies retained information in intermediate features by measuring mutual information between those features and the labels to be forgotten, offering a more comprehensive assessment of unlearning efficacy. Our experiments demonstrate that IDI effectively measures the degree of unlearning across various datasets and architectures, providing a reliable tool for evaluating strong unlearning in DNNs.

Extended version — https://arxiv.org/abs/2405.17878

1 Introduction

Machine unlearning (MU) seeks to remove the impact of specific data samples from a trained model, addressing privacy issues such as "right to be forgotten" (Voigt and Von dem Bussche 2017). In addition to privacy, MU is also emerging as a tool to eliminate the influence of corrupted or outdated data used during training (Nguyen et al. 2022; Kurmanji et al. 2023). The most straightforward approach to MU is *exact unlearning*, where the model is retrained from scratch, excluding the data that need to be forgotten. Although this method ensures complete data removal, it is computationally expensive and not scalable (Aldaghri, Mahdavifar, and Beirami 2021; Bourtoule et al. 2021). Consequently, research has shifted towards *approximate unlearning*, which aims to replicate the effects of retraining in a more efficient manner.

The goal of MU is to create an unlearned model that is indistinguishable from a model retrained from scratch, referred to as strong unlearning. This objective has become particularly crucial with the rise of open-source models like Stable Diffusion (Rombach et al. 2022) and LLaMA (Touvron et al. 2023), which are widely used and fine-tuned by various users. For unlearning algorithms to be practically useful, they must be capable of fully eliminating traces of private data and preventing potential exploitation. While (ϵ, δ)-certified unlearning methods (Zhang et al. 2024b; Mu and Klabjan 2024) provide theoretical guarantees, they are impractical for large-scale models. As a result, most approximate unlearning methods rely on heuristic approaches, lacking formal guarantees. Thus, these methods must undergo empirical evaluation to demonstrate their effectiveness.

However, current evaluations, primarily based on black-box approaches such as membership inference attacks (MIA) (Shokri et al. 2017) and accuracy comparisons, focus on output similarity rather than internal model changes. Although these metrics may capture weak unlearning (Fan et al. 2024; Chundawat et al. 2023a), they may not be sufficient for assessing strong unlearning. In this work, we study whether relying solely on outputs can truly reflect complete influence removal, given that model outputs can be superficially adjusted (Kirichenko, Izmailov, and Wilson 2023).

Surprisingly, our experiments reveal that even minimal changes—modifying only the final layer while preserving all intermediate information—can satisfy black-box evaluation metrics, exposing their limitations in assessing strong unlearning. This raises critical concerns about whether current MU methods truly achieve information removal comparable to retraining from scratch.

Consequently, motivated by the Information Bottleneck principle (Tishby, Pereira, and Bialek 2000), we introduce the **information difference index (IDI)**, a novel white-box metric designed to quantify residual information in intermediate layers after unlearning. IDI measures the mutual information (Shannon 1948) between intermediate features and the forgetting labels, providing an interpretable value to assess the effectiveness of unlearning algorithms. IDI remains stable under the stochasticity of the unlearning process and is compatible with diverse model architectures. Moreover, IDI can be reliably estimated from a data subset, making it practical for large-scale unlearning settings.

[*]These authors contributed equally.
[†]Corresponding authors.

Through IDI, we observe that many recent MU methods, despite strong performance on black-box metrics, retain substantial information about the forgetting data in intermediate layers. To address this, we propose **COLapse-and-Align (COLA)**, a simple method that collapses representations associated with the forget set at the feature level to eliminate residual information, followed by re-aligning the retained features. Despite its simplicity, COLA consistently improves IDI scores while maintaining competitive performance across datasets such as CIFAR-10/100, and ImageNet-1K, and architectures (ResNet-18/50, ViT).

2 Problem Statement and Preliminaries

2.1 Problem Statement

Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ denote a training dataset comprising N image-label pairs (x_i, y_i). In a supervised learning setup, D is partitioned into two subsets: the *forget set* D_f, containing the data points to be removed, and the *retain set* $D_r = D \setminus D_f$, containing the data points to be preserved. The initial model θ_o, referred to as the **Original model**, is trained on the full dataset D using empirical risk minimization. The **Retrain model** θ_r is trained from scratch on only the retain set D_r. The **unlearned model** θ_u is obtained by applying a MU algorithm to the Original model θ_o, aiming to remove the influence of D_f. The goal of MU is for θ_u to closely approximate θ_r, ensuring the unlearned model behaves as though D_f had never been used in training, while preserving the training methodology across θ_o, θ_r, and θ_u.

Throughout the paper, within a given model θ, we define the **head** as the last few layers responsible for classification; typically one to three linear layers. The **encoder**, on the contrary, encompasses the remainder of the network, which usually consists of convolutional layers or transformer encoders. MU is often studied in the context of image classification (Shaik et al. 2023; Nguyen et al. 2022), where it is typically classified into two scenarios based on the nature of the forget set: **class-wise forgetting**, where all samples from a specific class are targeted, and **random data forgetting**, where samples are selected indiscriminately across all classes. In this work, we tackle both scenarios.

2.2 Preliminaries

Machine Unlearning (MU). Exact unlearning, which involves creating Retrain, guarantees the information removal from the forget set but is computationally expensive (Bourtoule et al. 2021; Yan et al. 2022). To address this, approximate unlearning methods have been developed, focusing on efficiency rather than strict theoretical guarantees. Specifically, strong unlearning, where the unlearned model is indistinguishable from Retrain, has been explored through the application of differential privacy (DP) (Dwork and Roth 2014) inspired techniques, which aim to achieve parameter-level indistinguishability (Dwork and Roth 2014; Neel, Roth, and Sharifi-Malvajerdi 2021; Sekhari et al. 2021). However, applying such techniques to neural networks remains challenging due to their vast number of parameters and non-convex loss landscapes (Qiao et al. 2024). As a result, recent studies typically assess the similarity of model outputs (i.e., predictions), using weak unlearning as a practical proxy for strong unlearning (Xu et al. 2023).

Although empirically ensuring strong unlearning is challenging, it remains essential for deploying unlearning algorithms in compliance with legal requirements such as the GDPR (Voigt and Von dem Bussche 2017) and the "right to be forgotten." This need is further amplified by the widespread use of open-source models like CLIP (Radford et al. 2021), Stable Diffusion (Rombach et al. 2022), and LLaMA (Touvron et al. 2023), where sensitive data may inadvertently persist and be exploited. Our work focuses on developing a robust empirical metric to evaluate unlearning algorithms, distinct from verification (Zhang et al. 2024a; Sommer et al. 2022), which assesses effectiveness through real-world attack scenarios.

Evaluation Criteria in MU. As the goal of MU is to remove the influence of specific data while preserving the others, the unlearning algorithms are typically evaluated on three criteria: *Efficacy*, *Accuracy*, and *Efficiency* (Hayes et al. 2024). Efficacy measures how closely the unlearned model approximates Retrain, which is key to unlearning quality. Accuracy ensures task performance remains intact after unlearning, while efficiency ensures the unlearning process is faster than retraining.

Accuracy and efficiency can be easily evaluated using existing metrics. Accuracy consists of three categories: *unlearning accuracy (UA)*, *remaining accuracy (RA)*, and *testing accuracy (TA)*. UA measures performance on $\mathcal{D}_f$ as $UA(\theta_{\mathbf{u}}) = 1 - \text{Acc}_{\mathcal{D}_f}(\theta_{\mathbf{u}})$, RA on $\mathcal{D}_r$ as $RA(\theta_{\mathbf{u}}) = \text{Acc}_{\mathcal{D}_r}(\theta_{\mathbf{u}})$, and TA measures generalization to unseen data as $TA(\theta_{\mathbf{u}}) = \text{Acc}_{\mathcal{D}_{test}}(\theta_{\mathbf{u}})$. Performance levels comparable to Retrain across these metrics indicate better unlearning.

In terms of efficiency, *runtime efficiency (RTE)* measures the time an algorithm takes to complete unlearning, with lower RTE indicating more efficient unlearning (Fan et al. 2024; Jia et al. 2023). However, assessing unlearning efficacy, or determining whether the unlearned model has fully removed the influence of specific data to the same extent as Retrain, remains a significant challenge in complex DNNs. The efficacy metrics are divided into two categories: *black-box* metrics, which focus solely on model outputs (i.e., predictions), and *white-box* metrics, which examine internal dynamics such as parameters, gradients, and features. While black-box metrics are typically used due to their convenience, no universally accepted standard exists, leaving room for more reliable assessment.

Black-box Efficacy Metrics. Variants of membership inference attacks (MIA) (Shokri et al. 2017) are among the most widely used black-box metrics for evaluating unlearning (Fan et al. 2024; Jia et al. 2023; Foster, Schoepf, and Brintrup 2024). MIA trains an auxiliary classifier to determine whether a given sample was part of the training set, with attack success rates on the forget set close to those for Retrain being preferred in unlearning. Recent studies often combine MIA with UA, RA, TA, and RTE to assess unlearning across efficacy, accuracy, and efficiency (Chen et al. 2023; Kim, Lee, and Woo 2024), a practice referred to as the 'full-stack' evaluation (Jia et al. 2023; Fan et al. 2024).

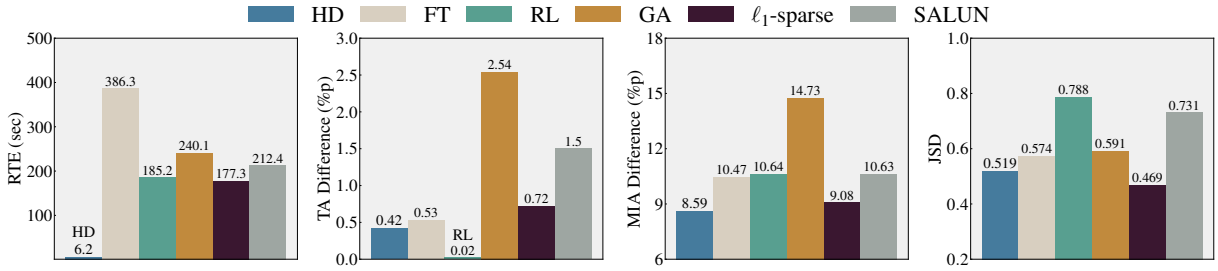


Figure 1: Performance of six methods on (CIFAR-10, ResNet-18), evaluated in efficiency (RTE), accuracy (TA), and efficacy (MIA, JSD). For TA, MIA, and JSD, lower differences from Retrain are preferred, indicating closer similarity to Retrain.

Other metrics, such as Jensen-Shannon divergence (JSD), and ZRF (Chundawat et al. 2023a; Poppi et al. 2024) compare the output logits between the unlearned model and Retrain (or a random model for ZRF). Additionally, time-based metrics like Anamnesis Index (AIN) (Chundawat et al. 2023b; Tarun et al. 2023a) and relearn time (RT) (Tarun et al. 2023b) track how long the model takes to regain performance on the forget set. While convenient, black-box metrics overlook internal behaviors and cannot verify strong unlearning by ensuring forgetting data's influence is fully removed. Their limitations are further discussed in Section 3.

White-box Efficacy Metrics. In contrast, white-box metrics offer deeper insights by analyzing internal model dynamics. Previous studies have measured parameter-wise distances (e.g., ℓ_2-distance, KL-divergence) between the unlearned model and Retrain (Golatkar, Achille, and Soatto 2020; Wu, Dobriban, and Davidson 2020). However, this approach is computationally expensive and unreliable due to training randomness (Hayes et al. 2024; Goel et al. 2022). Becker and Liebig (2022) proposed a Fisher information based metric, but their results were inconsistent with theoretical intuition. Graves, Nagisetty, and Ganesh (2021) applied model inversion attacks to reconstruct images from the forget set, but their method relies on visual comparisons.

Although robust white-box metrics are currently lacking and challenging to develop, they are crucial for validating approximate methods that lack formal guarantees. Without them, these algorithms cannot be trusted in privacy-sensitive applications that demand a high level of confidence in information removal. To address this critical need, our work proposes a reliable and practical white-box metric.

3 Rethinking the Evaluation of Unlearning

3.1 Challenging Black-box Metrics

In this section, we reveal the limitations of commonly used black-box efficacy metrics by applying a simple unlearning technique to a single-class forgetting task. We show that these metrics can misrepresent unlearning efficacy, even when the model's output closely resembles that of Retrain.

Inspired by the teacher-student framework, our strategy, termed **head distillation (HD)**, employs logit distillation from Original $\theta_{\mathbf{o}}$. The unlearned model $\theta_{\mathbf{u}}$ is initialized from $\theta_{\mathbf{o}}$ with the encoder frozen and only the head trainable. During unlearning, the head is finetuned on training dataset $\mathcal{D}$ using KL-divergence loss (Hinton, Vinyals, and Dean 2014) to match a masked version of $\theta_{\mathbf{o}}$'s output, where the logit for the forgetting class is set to negative infinity. This approach enables $\theta_{\mathbf{u}}$ to mimic a pseudo-retrained model, as the masked logits closely resemble those of Retrain. By aligning output behavior, HD approximates the intended unlearning effect. Details in Appendix C.1 of (Jeon et al. 2025).

We evaluated HD on CIFAR-10 (Krizhevsky 2009) using ResNet-18 (He et al. 2016), where the head is only a single linear layer. For *efficacy*, we used membership inference attack (MIA) and Jensen-Shannon divergence (JSD). For *accuracy* and *efficiency*, we measured unlearning accuracy (UA), testing accuracy (TA), and run-time efficiency (RTE). We compared HD with recent methods, including FT, RL (Golatkar, Achille, and Soatto 2020), GA (Thudi et al. 2022), ℓ_1-sparse (Jia et al. 2023), and SALUN (Fan et al. 2024). Details on metrics and baselines can be found in Appendices C.2 and C.3 (Jeon et al. 2025).

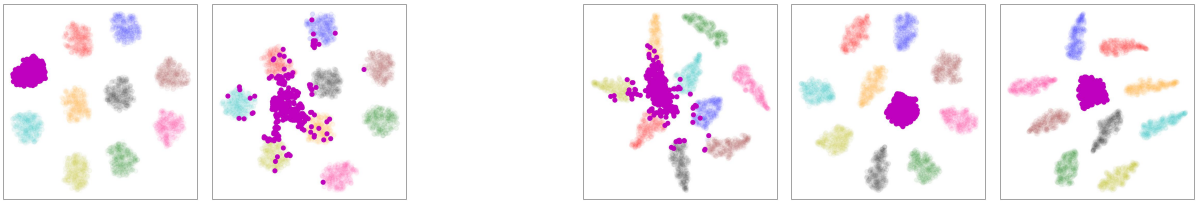
Figure 1 shows the experimental results. Despite its simplicity, HD outperforms all other methods in MIA and ranking second in JSD. HD achieves this performance in just 6.2 seconds, approximately 30 to 60 times faster than competing methods. Additionally, HD maintains comparable testing accuracy (TA), effectively preserving task performance. All methods achieved perfect unlearning accuracy (100% UA), which is omitted from Figure 1. Notably, HD's strong performance generalize to multi-class and random data forgetting, as shown in Appendix D.1 of (Jeon et al. 2025).

The results indicate that HD performs exceptionally well across all black-box metrics. However, its validity as a MU algorithm requires scrutiny. The primary issue is that HD closely resembles Original $\theta_{\mathbf{o}}$, with changes limited to the single-layer head, while the encoder remains identical to $\theta_{\mathbf{o}}$. Consequently, all intermediate features related to the forget set are perfectly retained. This raises a critical question:

> *Do black-box metrics truly capture the unlearning quality, or are they misled by superficial changes while deeper information persists?*

3.2 Residual Information of Forgetting Data

To address the above question, we analyze recent unlearning methods to assess whether they internally remove informa-



(a) Original / HD (b) Retrain (c) GA (d) RL (e) SALUN

Figure 2: t-SNE visualizations of encoder outputs for Original, Retrain, and unlearned models from three MU methods (GA, RL, SALUN) on single-class forgetting with (CIFAR-10, ResNet-18). In each t-SNE plot, features of the forgetting class are represented in purple. Original and HD have identical feature distribution as they share the same encoder.

tion from the forget set, despite their strong performance on black-box metrics. Note that analyses use the same experimental setup described in Section 3.1.

We begin with a qualitative analysis using t-SNE (van der Maaten and Hinton 2008) visualizations of intermediate features from model encoders to compare Retrain and Original, and to examine internal behaviors of unlearning methods (Figure 2). In Figure 2b, the forgetting class (in purple) shows a scattered distribution in Retrain, indicating difficulty in forming coherent representations. This scattering reflects an desirable outcome of strong unlearning, suggesting that the model has successfully 'forgotten' how to encode meaningful semantic information from the forget set.

Notably, while the features from GA (Thudi et al. 2022) appear scattered in a manner similar to Retrain, the features from RL (Golatkar, Achille, and Soatto 2020) and SALUN (Fan et al. 2024) closely resemble those of Original. In addition, HD, which shares the same encoder as Original, shows identical t-SNE results. These findings indicate that several unlearned models still retain a significant capacity to recognize the forgetting class, unlike Retrain.

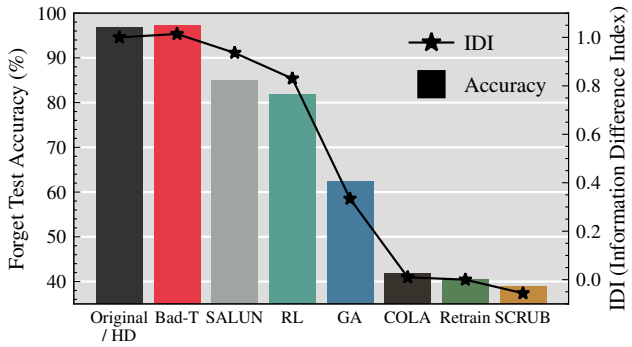


Figure 3: Forget test accuracy and IDI (our metric in Section 4.3) for Original, Retrain, and MU methods (including COLA, our method in Section 5.1) after head retraining with fixed unlearned encoders using 2% of $\mathcal{D}$ in (CIFAR-10, ResNet-18). IDI aligns with the recovered accuracy.

To further examine the residual influence in unlearned models, we conducted a follow-up experiment inspired by time-based metrics (*e.g.*, Chundawat et al. (2023b)). We test whether unlearned encoders can recover forgotten information using minimal data. Specifically, we replaced the heads of all models, including Retrain and Original, with randomly initialized ones. The encoders were then frozen, and new heads were trained on $\mathcal{D}'$, a small subset (only 2% of the total) of $\mathcal{D}$ sampled at random.

After training, we evaluated the accuracy of the new models on the forget test data. Surprisingly, as shown in Figure 3, while the retrained head of Retrain achieves no more than 41% accuracy, the heads from certain methods, like Bad-T, SALUN, and RL exhibit over 82% accuracy. Notably, the high accuracy observed in SALUN and RL aligns with their clustered t-SNE patterns in Figure 2.

The above results show that unlearned models across various MU algorithms retain substantial residual influence from the forget set, indicating incomplete unlearning. Critically, standard black-box metrics fail to capture these internal traces. If such metrics cannot ensure strong unlearning, the reliability of approximate unlearning algorithms, which often lack theoretical guarantees, becomes questionable in real world applications. Therefore, developing practical white box approaches that consider internal model behaviors is essential to achieving the fundamental goal of unlearning.

4 An Information Theoretic Metric

Black-box metrics often miss residual information in intermediate layers, as shown in Section 3. To capture this, we measure residual information in intermediate features using mutual information and introduce IDI, a white-box metric that evaluates unlearning beyond outputs.

4.1 Quantifying Residual Information

To quantify the relationship between intermediate features and data labels, we utilize Shannon's mutual information (MI), a robust measure that captures variable dependencies across dimensional complexities. For an input $\mathbf{X}$, let $\mathbf{Z}_\ell^{(\mathbf{u})}$ and $\mathbf{Z}_\ell^{(\mathbf{r})}$ denote the features from the ℓ-th layer of the total L-layer encoder in the unlearned model and Retrain, respectively. Let Y be a binary label indicating whether $\mathbf{X}$ belongs to the forget set ($Y = 1$) or not ($Y = 0$). We compute MI $I(\mathbf{Z}_\ell; Y)$ across each layer from 1 to L, to determine whether intermediate features retain information about the forget set. For estimation, we adopt the InfoNCE loss (Oord, Li, and Vinyals 2018), a robust method widely used in deep MI estimation (Radford et al. 2021; Jia et al. 2021).

Given a batch $\mathcal{B} = \{(U^{(k)}, V^{(k)}) : 1 \le k \le K\}$, sampled from a joint distribution $P_{U,V}$, where $U \in \mathcal{U}$ and $V \in \mathcal{V}$ be random variables. The InfoNCE loss (Poole et al. 2019) is defined as:

$$\mathcal{L}_{\text{NCE}} = \frac{1}{K}\sum_{k=1}^{K} \log \frac{\exp(f_\nu(U^{(k)})^\top g_\eta(V^{(k)}))}{\frac{1}{K}\sum_{k'=1}^{K}\exp(f_\nu(U^{(k)})^\top g_\eta(V^{(k')}))},$$

where $f_\nu : \mathcal{U} \to \mathbb{R}^d$ and $g_\eta : \mathcal{V} \to \mathbb{R}^d$ are critic functions, with an output embedding dimension d, parameterized by neural networks with parameters ν and η. This neural network parameterization, inspired by Radford et al. (2021), effectively captures complex relationships in contrastive learning through flexible and expressive modeling of the joint distributions of U and V.

The InfoNCE loss serves as a lower bound on the MI between U and V. In fact, the maximum value of the InfoNCE loss, when using the joint critic functions, equals the mutual information:

$$I(U;V) = \max_{\nu,\eta} \mathcal{L}_{\text{NCE}}(\mathcal{B}, \nu, \eta).$$

By maximizing this loss over parameters ν and η through neural networks, we effectively capture data structure and accurately quantify shared information between U and V.

To estimate mutual information (MI) at each layer, we define separate critic functions for every layer: f_{ν_ℓ} and g_{η_ℓ}, where $\ell \in \{1, \dots, L\}$ denotes the layer index. The critic g_{η_ℓ} models the binary variable Y as two trainable d-dimensional vectors, $g_{\eta_\ell}(0)$ and $g_{\eta_\ell}(1)$, selecting the appropriate one based on the value of Y. In parallel, f_{ν_ℓ} maps intermediate features $\mathbf{Z}_\ell$ from the ℓ-th encoder layer to a shared d-dimensional embedding space. The parameters ν_ℓ define the weights and biases of this neural network.

The complexity of f_{ν_ℓ} depends on the layer depth: in earlier layers, it processes raw, less interpretable features, requiring intricate design to capture the relationship between $\mathbf{Z}_\ell$ and Y, while later layers with structured features allow more direct mapping. This design enables the accurate estimation of $I(\mathbf{Z}_\ell; Y)$, capturing the dependency between features and labels at different depths. For details on f_{ν_ℓ} and g_{η_ℓ}, refer to Appendix B of (Jeon et al. 2025).

For model-agnostic design, we construct f_{ν_ℓ} by reusing the network layers from $\ell + 1$ to L. This approach allows us to approximate the mutual information between the output and intermediate features at layer ℓ without requiring network redesign for each layer, maintaining flexibility and scalability. To ensure dimensional compatibility between f and g, we introduce an additional linear projection layer so that $f_{\nu_\ell}(\mathbf{Z}_\ell)$ outputs a d-dimensional feature.

During optimization, we freeze the parameters of the model up to the ℓ-th layer and reuse the subsequent layers, from $\ell + 1$ to L, as f_{ν_ℓ}. These layers, together with the projection layer, are randomly initialized and trained to optimize the InfoNCE objective. Both the retain and forget sets are used to provide representations for $Y = 0$ and $Y = 1$, ensuring that information from both outcomes is captured for mutual information estimation.

This approach enables f_{ν_ℓ} to effectively exploit intermediate features $\mathbf{Z}_\ell$ to classify Y, providing deeper insights into the model's internal information processing at each layer. It also reveals the model's capacity to extract and utilize relevant information for distinguishing between output labels, offering a clearer understanding of the information dynamics across the network.

4.2 Residual Information in Unlearned Models

We begin by plotting the estimated MI between the intermediate layers and the binary label indicating whether the data belong to the forget set, as shown in Figure 4b. As expected, MI decreases across layers, aligning with the Information Bottleneck principle (Tishby, Pereira, and Bialek 2000). This figure also reveals the internal behaviors of unlearned models that black-box assessments fail to capture.

In particular, SCRUB and ℓ_1-sparse, which approximate the MI levels of Retrain, are more likely to achieve the MU objective at the feature level across both ResNet architectures. Their lower MI suggests that their encoders, like Retrain, struggle to differentiate between the forget set and the retain set. Conversely, SALUN and RL show MI curves that are close to that of Original, indicating the opposite. Note that HD produces the identical curve as Original, as its encoder remains unchanged. We observe similar patterns in CIFAR-100 and ImageNet-1K, as well as in ViT. Additionally, extending our experiment to multi-class forgetting tasks (*e.g.*, 20 classes on CIFAR-100) reveals more pronounced MI differences between Retrain and Original. See Appendix E.1 of Jeon et al. (2025) for further results.

4.3 Information Difference Index (IDI)

Motivated from the above experiment, we define the **information difference (ID)** of $\theta_{\mathbf{u}}$ as the MI difference across intermediate layers between the unlearned model and Retrain, calculated as:

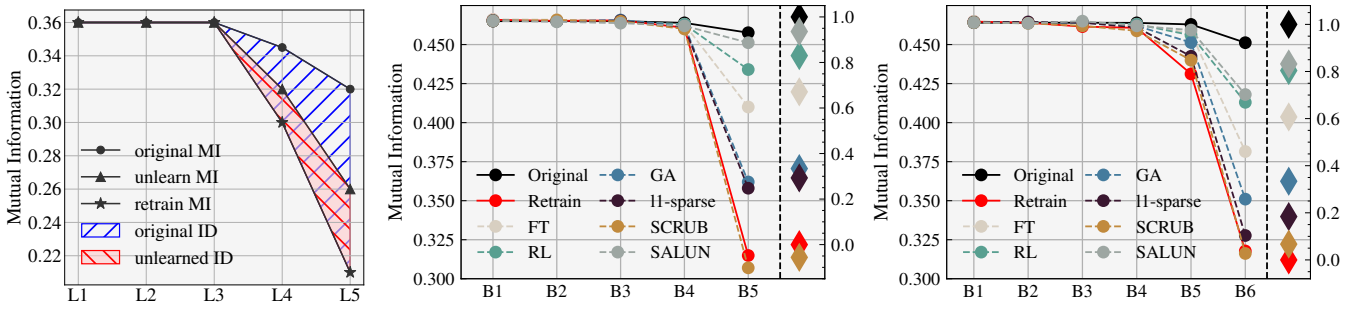
$$\mathbf{ID}(\theta_{\mathbf{u}}) = \sum_{\ell=1}^{L} \big(I(\mathbf{Z}_\ell^{(\mathbf{u})}; Y) - I(\mathbf{Z}_\ell^{(\mathbf{r})}; Y)\big). \tag{1}$$

ID of $\theta_{\mathbf{u}}$ shows the extent of information retention through ensuing layers of the unlearned encoder. To provide a normalized measure, we introduce the **information difference index (IDI)**:

$$\mathbf{IDI}(\theta_{\mathbf{u}}) = \frac{\mathbf{ID}(\theta_{\mathbf{u}})}{\mathbf{ID}(\theta_{\mathbf{0}})} = \frac{\sum_{\ell=1}^{L}\big(I(\mathbf{Z}_\ell^{(\mathbf{u})}; Y) - I(\mathbf{Z}_\ell^{(\mathbf{r})}; Y)\big)}{\sum_{\ell=1}^{L}\big(I(\mathbf{Z}_\ell^{(\mathbf{o})}; Y) - I(\mathbf{Z}_\ell^{(\mathbf{r})}; Y)\big)}, \tag{2}$$

where $\mathbf{Z}_\ell^{(\mathbf{o})}$ is the output of the ℓ-th layer of Original encoder. Figure 4a illustrates IDI, which is conceptually the ratio of the areas between MI curves.

However, computing MI for all L layers can be expensive. In practice, we compute MI from the last n layers (*i.e.*, later blocks), where $n \ll L$, as earlier layers show negligible differences between Retrain and Original (Figure 4b). To reduce overhead, we reuse the original network structure for MI estimation (see Figure 5). With this setup, computing IDI on CIFAR-100 with ResNet-18 takes under 5 minutes. Detailed cost analysis in Appendix E.2 of (Jeon et al. 2025).



(a) Illustraion of IDI

(b) MI curves and IDI in single-class forgetting task

Figure 4: (a) Conceptual illustration of IDI. Curves show estimated mutual information $I(\mathbf{Z}_\ell; Y)$ for Original (●), unlearned (▲), and Retrain (★). IDI is the ratio $\frac{\mathbf{ID}(\theta_u)}{\mathbf{ID}(\theta_o)}$, corresponding to the red area divided by the blue area. (b) MI curves and IDI values for Original, Retrain, and unlearned models (FT, RL, GA, ℓ_1-sparse, SCRUB, SALUN) on CIFAR-10 across ResNet-18 (left) and ResNet-50 (right) blocks, averaged over five trials. See Appendix D.2 of (Jeon et al. 2025) for standard deviations.

	CIFAR-10 (single class)						ImageNet-1K (five classes)					
Methods	UA	RA	TA	MIA	**IDI**	RTE (min)	UA	RA	TA	MIA	**IDI**	RTE (min)
Retrain	100.0	100.0	95.64	10.64	0.0	154.56	100.0	88.80	75.88	9.41	0.0	2661.90
HD	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	95.22$_{\pm 0.07}$	2.05$_{\pm 0.11}$	1.000$_{\pm 0.0}$	**0.10**$_{\pm 0.01}$	**100.0**$_{\pm 0.0}$	87.94$_{\pm 0.16}$	75.60$_{\pm 0.07}$	7.12$_{\pm 0.12}$	1.000$_{\pm 0.0}$	**4.75**$_{\pm 0.03}$
FT	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	95.12$_{\pm 0.09}$	0.17$_{\pm 0.05}$	0.671$_{\pm 0.008}$	6.44$_{\pm 0.07}$	**100.0**$_{\pm 0.0}$	**88.52**$_{\pm 0.0}$	76.16$_{\pm 0.01}$	8.24$_{\pm 1.23}$	0.102$_{\pm 0.026}$	140.04$_{\pm 1.42}$
RL	99.93$_{\pm 0.01}$	**100.0**$_{\pm 0.0}$	**95.66**$_{\pm 0.05}$	0.0$_{\pm 0.0}$	0.830$_{\pm 0.005}$	3.09$_{\pm 0.03}$	99.96$_{\pm 0.03}$	86.46$_{\pm 0.07}$	75.23$_{\pm 0.01}$	0.23$_{\pm 0.01}$	1.002$_{\pm 0.007}$	200.73$_{\pm 1.87}$
GA	**100.0**$_{\pm 0.0}$	99.06$_{\pm 0.25}$	93.10$_{\pm 0.50}$	25.37$_{\pm 3.24}$	0.334$_{\pm 0.014}$	4.00$_{\pm 0.08}$	**100.0**$_{\pm 0.0}$	80.77$_{\pm 0.22}$	71.49$_{\pm 0.10}$	4.20$_{\pm 0.46}$	0.328$_{\pm 0.023}$	212.14$_{\pm 2.61}$
Bad-T	99.90$_{\pm 0.14}$	99.99$_{\pm 0.0}$	94.99$_{\pm 0.12}$	68.17$_{\pm 42.80}$	1.014$_{\pm 0.004}$	4.64$_{\pm 0.05}$	98.01$_{\pm 0.02}$	84.03$_{\pm 0.03}$	73.42$_{\pm 0.03}$	69.13$_{\pm 12.57}$	1.152$_{\pm 0.011}$	211.52$_{\pm 0.96}$
EU-5	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	95.25$_{\pm 0.02}$	0.06$_{\pm 0.03}$	0.528$_{\pm 0.005}$	1.54$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	79.62$_{\pm 0.0}$	71.22$_{\pm 0.13}$	13.33$_{\pm 1.53}$	0.183$_{\pm 0.028}$	193.38$_{\pm 0.78}$
CF-5	98.13$_{\pm 1.39}$	**100.0**$_{\pm 0.0}$	95.54$_{\pm 0.09}$	0.0$_{\pm 0.0}$	0.675$_{\pm 0.027}$	1.57$_{\pm 0.03}$	**100.0**$_{\pm 0.0}$	84.31$_{\pm 0.08}$	74.16$_{\pm 0.06}$	10.21$_{\pm 5.33}$	0.701$_{\pm 0.014}$	81.53$_{\pm 0.56}$
EU-10	**100.0**$_{\pm 0.0}$	99.50$_{\pm 0.02}$	93.61$_{\pm 0.08}$	15.24$_{\pm 1.08}$	−0.349$_{\pm 0.019}$	2.42$_{\pm 0.11}$	**100.0**$_{\pm 0.0}$	71.84$_{\pm 0.03}$	65.78$_{\pm 0.02}$	16.65$_{\pm 1.91}$	−0.051$_{\pm 0.021}$	193.79$_{\pm 0.47}$
CF-10	**100.0**$_{\pm 0.0}$	99.98$_{\pm 0.0}$	94.95$_{\pm 0.05}$	**11.61**$_{\pm 0.91}$	−0.060$_{\pm 0.017}$	2.31$_{\pm 0.03}$	**100.0**$_{\pm 0.0}$	80.87$_{\pm 0.04}$	72.34$_{\pm 0.08}$	13.99$_{\pm 5.41}$	0.608$_{\pm 0.012}$	82.29$_{\pm 0.34}$
SCRUB	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	95.37$_{\pm 0.04}$	19.73$_{\pm 1.92}$	−0.056$_{\pm 0.008}$	3.49$_{\pm 0.02}$	99.28$_{\pm 0.07}$	88.39$_{\pm 0.04}$	76.51$_{\pm 0.03}$	7.42$_{\pm 0.51}$	0.517$_{\pm 0.011}$	426.04$_{\pm 2.98}$
SALUN	99.99$_{\pm 0.01}$	**100.0**$_{\pm 0.0}$	95.42$_{\pm 0.12}$	0.01$_{\pm 0.01}$	0.936$_{\pm 0.012}$	3.54$_{\pm 0.11}$	89.67$_{\pm 0.27}$	86.25$_{\pm 0.15}$	75.54$_{\pm 0.10}$	0.50$_{\pm 0.09}$	0.343$_{\pm 0.017}$	793.82$_{\pm 3.32}$
ℓ_1-sparse	**100.0**$_{\pm 0.0}$	99.93$_{\pm 0.02}$	94.90$_{\pm 0.10}$	1.56$_{\pm 0.09}$	0.293$_{\pm 0.012}$	2.96$_{\pm 0.03}$	97.57$_{\pm 0.61}$	85.33$_{\pm 0.07}$	74.77$_{\pm 0.03}$	8.84$_{\pm 1.39}$	0.239$_{\pm 0.031}$	226.74$_{\pm 1.35}$
COLA	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	95.36$_{\pm 0.06}$	12.64$_{\pm 0.92}$	**0.010**$_{\pm 0.006}$	4.91$_{\pm 0.04}$	**100.0**$_{\pm 0.0}$	87.93$_{\pm 0.05}$	**76.15**$_{\pm 0.04}$	**9.95**$_{\pm 1.21}$	**0.040**$_{\pm 0.042}$	171.44$_{\pm 0.75}$

Table 1: Performance summary of MU methods (including COLA and 14 other baselines) for class-wise forgetting task on (CIFAR-10, ResNet-18) and (ImageNet-1K, ResNet-50). A better performance of an MU method corresponds to a smaller performance gap with Retrain (except RTE), with the top method in **bold** and the second best underlined.

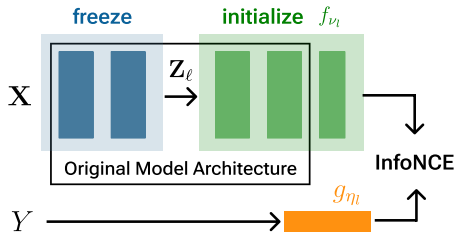


Figure 5: Illustration of estimating MI using InfoNCE. f_{ν_ℓ} represents a trainable network to capture features from $\mathbf{Z}_\ell$, while g_{η_ℓ} handles the binary input Y.

IDI quantifies the information gap between the unlearned model and Retrain. An IDI of 0 indicates complete removal of forget-set information, achieving indistinguishability from Retrain. In contrast, an IDI of 1 indicates that the encoder retains all the information found in Original. Interestingly, a negative IDI value, termed *over-unlearning*, occurs when the model removes more information than Retrain. We also demonstrate IDI for random data forgetting in Appendix A.1 of (Jeon et al. 2025)

5 Experiments

5.1 COLlapse and Align (COLA) Approach

As discussed in both Section 3 and 4.2, several unlearned models retain residual information in intermediate layers even when their outputs match Retrain. To resolve this, we introduce **COLlapse and Align (COLA)**, a two-step framework consisting of a *collapse phase* and an *alignment phase* that removes residual information at the feature level.

During the *collapse phase*, COLA eliminates feature-level information by applying supervised contrastive loss (Khosla et al. 2020) to encoder outputs. Rather than dispersing features from the forget set, which could harm model performance, COLA applies the loss to the retain set, promoting tight intra-class clustering. As these clusters shrink, features from the forget set are forced to collapse into the clusters of the retain set, achieving catastrophic forgetting. After feature collapsing, the *alignment phase* optimizes the entire model using cross-entropy loss on the retain set to align the encoder and head. For an intuitive illustration of COLA, as well as COLA+, a method tailored for random data forgetting, along with their detailed objective formulations, refer to Appendix C.7 of (Jeon et al. 2025).

5.2 Evaluation of Unlearning Methods with IDI

We demonstrate the utility of IDI as a valuable efficacy metric and highlight the strong performance of COLA and its variant COLA+ through extensive experiments. Our experiments cover three datasets: CIFAR-10, CIFAR-100 (Krizhevsky 2009), and ImageNet-1K (Deng et al. 2009), and three model architectures: ResNet-18, ResNet-50 (He et al. 2016), and ViT (Dosovitskiy et al. 2021). For simplicity, we approximate IDI using the features from blocks rather than every layer in ResNet and ViT. See Appendix C of (Jeon et al. 2025) for experimental details.

CIFAR-10 (500 samples per class)						
Methods	UA	RA	TA	MIA	**IDI**	RTE
Retrain	3.94	100.0	95.26	75.12	0.0	152.87
HD	$\underline{3.64}_{\pm 1.66}$	$97.93_{\pm 1.38}$	$92.80_{\pm 1.18}$	$\underline{77.47}_{\pm 4.09}$	$1.000_{\pm 0.0}$	**0.30**
FT	$5.03_{\pm 0.40}$	$98.95_{\pm 0.21}$	$92.94_{\pm 0.26}$	$83.52_{\pm 0.58}$	$\underline{-0.069}_{\pm 0.013}$	8.11
RL	$4.77_{\pm 0.27}$	$\mathbf{99.92}_{\pm 0.0}$	$\mathbf{93.54}_{\pm 0.04}$	$22.47_{\pm 1.19}$	$0.084_{\pm 0.030}$	2.75
GA	$2.86_{\pm 0.76}$	$98.37_{\pm 0.71}$	$91.90_{\pm 0.70}$	$85.49_{\pm 2.17}$	$0.924_{\pm 0.028}$	4.31
Bad-T	$5.47_{\pm 1.05}$	$\underline{99.87}_{\pm 0.05}$	$91.51_{\pm 0.61}$	$39.53_{\pm 3.43}$	$0.939_{\pm 0.053}$	4.78
EU-10	$3.16_{\pm 0.19}$	$98.68_{\pm 0.08}$	$93.07_{\pm 0.12}$	$83.40_{\pm 0.21}$	$-0.110_{\pm 0.013}$	2.13
CF-10	$2.71_{\pm 0.24}$	$99.11_{\pm 0.06}$	$\underline{93.47}_{\pm 0.15}$	$84.33_{\pm 0.05}$	$0.219_{\pm 0.029}$	$\underline{2.10}$
SCRUB	$4.31_{\pm 1.50}$	$96.21_{\pm 1.70}$	$88.83_{\pm 1.86}$	$37.88_{\pm 7.65}$	$0.322_{\pm 0.016}$	3.37
SALUN	$2.74_{\pm 0.30}$	$97.77_{\pm 0.04}$	$91.68_{\pm 0.44}$	$83.52_{\pm 2.20}$	$0.861_{\pm 0.012}$	5.69
ℓ_1-sparse	$5.47_{\pm 0.22}$	$96.66_{\pm 0.07}$	$91.31_{\pm 0.25}$	$\mathbf{77.12}_{\pm 0.21}$	$-0.157_{\pm 0.026}$	3.03
COLA+	$\mathbf{3.90}_{\pm 0.08}$	$99.24_{\pm 0.17}$	$93.23_{\pm 0.09}$	$83.48_{\pm 0.10}$	$\mathbf{0.024}_{\pm 0.010}$	7.80

Table 2: Performance summary for random data forgetting on (CIFAR-10, ResNet-18), with the top method in **bold** and the second best underlined. RTE is reported in minutes.

Table 1 shows the experimental results on CIFAR-10 and ImageNet-1K in class-forgetting tasks. At first glance, excluding the IDI column, several methods show similar accuracy (UA, RA, TA) but greater deviations in efficacy (MIA) and efficiency (RTE). This suggests that previous unlearning studies likely ranked MU methods based on MIA and RTE. However, as discussed earlier, relying solely on black-box metrics can be misleading, as they fail to account for residual information. Indeed, some methods show strong MIA performance but fail to remove forget data from intermediate layers, as reflected by high IDI values. For instance, CF-5 on ImageNet-1K achieves a favorable MIA value (10.21) close to Retrain (9.41) in the shortest time (81.53 min), yet its IDI (0.701) shows significant retention of forget data. Similarly, EU-5 on CIFAR-10, which appears highly efficient (1.54 min), presents a high IDI (0.528), suggesting that its efficiency stems from incomplete unlearning. The discrepancy between black-box metrics (MIA, JSD) and IDI is similarly observed in random data forgetting, as shown in Table 2, particularly for methods like SALUN. By incorporating IDI alongside existing metrics, we gain a more comprehensive and insightful evaluation of MU methods.

5.3 Discussions

IDI as a Real-World Efficacy Metric. Accuracy metrics (UA, RA, TA) and efficacy metrics (MIA, JSD), commonly used in recent unlearning studies, require the presence of Retrain as a gold standard to compare model outputs. While this approach is crucial for advancing MU methods in controlled experimental settings, where the field of unlearning for DNNs is still in its infancy, it becomes impractical in real-world applications where Retrain is unavailable. Similar to current black-box metrics, the original formulation of IDI (see Equations 1 and 2) uses Retrain as a reference to assess unlearning efficacy. However, IDI allows for flexibility by using any available unlearned model as the reference. Although the absence of Retrain changes the interpretation of IDI (*i.e.*, an IDI of zero means complete unlearning as Retrain), it still provides valuable insights relative to the chosen reference. This adaptability enhances the IDI's practicality, making it useful for evaluating unlearned models even in real-world scenarios. A detailed explanation and examples are provided in Appendix E.3 of (Jeon et al. 2025).

	CIFAR-10			CIFAR-100		
Methods	Activation	Gradient	**IDI**	Activation	Gradient	**IDI**
Original	$99.98_{\pm 0.03}$	$100.0_{\pm 0.0}$	1.00	$53.13_{\pm 2.88}$	$61.34_{\pm 3.23}$	1.000
Retrain	$94.89_{\pm 1.07}$	$95.13_{\pm 1.12}$	0.000	$52.87_{\pm 6.15}$	$59.12_{\pm 4.12}$	0.000
Random	$52.89_{\pm 41.03}$	$45.23_{\pm 23.04}$	$-1.281_{\pm 0.02}$	$53.20_{\pm 5.15}$	$47.12_{\pm 7.21}$	$-2.955_{\pm 0.05}$
RL	$100.0_{\pm 0.0}$	$99.98_{\pm 0.01}$	$0.830_{\pm 0.01}$	$93.20_{\pm 3.53}$	$95.30_{\pm 0.82}$	$0.467_{\pm 0.01}$
GA	$97.07_{\pm 0.35}$	$96.01_{\pm 0.13}$	$0.334_{\pm 0.01}$	$97.44_{\pm 2.12}$	$82.44_{\pm 0.95}$	$0.392_{\pm 0.02}$
EU-10	$86.13_{\pm 4.78}$	$89.42_{\pm 2.32}$	$-0.349_{\pm 0.02}$	$64.41_{\pm 1.65}$	$72.13_{\pm 4.13}$	$-0.221_{\pm 0.01}$
CF-10	$97.99_{\pm 0.38}$	$98.33_{\pm 0.23}$	$-0.060_{\pm 0.02}$	$21.62_{\pm 0.61}$	$23.15_{\pm 1.23}$	$0.175_{\pm 0.04}$
SCRUB	$99.43_{\pm 0.09}$	$99.15_{\pm 0.05}$	$-0.056_{\pm 0.00}$	$46.44_{\pm 1.28}$	$62.31_{\pm 1.73}$	$0.339_{\pm 0.07}$
COLA	$92.26_{\pm 0.08}$	$93.12_{\pm 0.11}$	$0.010_{\pm 0.00}$	$61.08_{\pm 0.23}$	$65.24_{\pm 0.43}$	$-0.037_{\pm 0.00}$

Table 3: Performance of MU methods on white-box MIAs (Activation, Gradient) and IDI for single-class forgetting on ResNet-18. MIA values represent the attack success rate (%) for distinguishing forgetting samples. "Random" refers to a model randomly initialized without prior training.

IDI compare to White-Box MIA. While black-box MIA, adapted from privacy studies, is widely used as an evaluation tool in unlearning literature, we explore the potential of white-box MIA, which has not traditionally been employed for this purpose, and compare it with IDI. Specifically, we evaluate two white-box MIA methods: one leveraging model activations and another utilizing gradients (Nasr, Shokri, and Houmansadr 2019). Table 3 presents the results of white-box MIA and IDI in single-class forgetting scenarios. White-box MIA delivers consistent results on CIFAR-10 but becomes unstable as the dataset scales to CIFAR-100, with significant variability in MIA values across algorithms. This instability is further highlighted with a randomly initialized model, which produces MIA values comparable to Retrain despite no actual training. In contrast, IDI provides stable and interpretable results, yielding strongly negative values for randomly initialized models, accurately reflecting their lack of residual information. This underscores IDI's reliability as a robust and interpretable metric for unlearning evaluation.

6 Conclusion

Black-box metrics fail to capture residual information in intermediate representations, limiting their ability to assess strong unlearning. We introduce the Information Difference Index (IDI), a white-box metric that quantifies retained information at the feature level. Experiments across datasets and architectures show that IDI provides a reliable evaluation of unlearning quality. We further propose COLA, an unlearning method that collapses and realigns representations to remove residual information directly.

Acknowledgements

This work was supported in part by IITP grants funded by the Korea government (MSIT) (No. RS-2024-00457882 (AI Research Hub Project), RS-2022-II220077, RS-2022-II220113, RS-2022-II220959, RS-2022-II220871, RS-2021-II211343 (SNU AI), RS-2025-25442338 (AI Star Fellowship-SNU)), a NRF grant funded by the Korea government (MSIT) (No. RS-2025-23525649), a grant (No. RS-2025-25453780) funded By MOTIE, a grant of Korean ARPA-H Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (RS-2025-25424639), and the BK21 FOUR program, SNU in 2025.

For completeness, we list the full institutional affiliations. Dongjae Jeon is with the Department of Computer Science at Yonsei University. Wonje Jeung and Albert No are with the Department of Artificial Intelligence at Yonsei University. Taeheon Kim and Jonghyun Choi are with the Department of Electrical and Computer Engineering at Seoul National University, and Jonghyun Choi is additionally affiliated with IPAI and ASRI at Seoul National University.

References

Aldaghri, N.; Mahdavifar, H.; and Beirami, A. 2021. Coded machine unlearning. *IEEE Access*.

Becker, A.; and Liebig, T. 2022. Evaluating machine unlearning via epistemic uncertainty. *arXiv preprint arXiv:2208.10836*.

Bourtoule, L.; Chandrasekaran, V.; Choquette-Choo, C. A.; Jia, H.; Travers, A.; Zhang, B.; Lie, D.; and Papernot, N. 2021. Machine unlearning. In *S&P*.

Chen, M.; Gao, W.; Liu, G.; Peng, K.; and Wang, C. 2023. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In *CVPR*.

Chundawat, V. S.; Tarun, A. K.; Mandal, M.; and Kankanhalli, M. 2023a. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *AAAI*.

Chundawat, V. S.; Tarun, A. K.; Mandal, M.; and Kankanhalli, M. 2023b. Zero-shot machine unlearning. *IEEE TIFS*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Dwork, C.; and Roth, A. 2014. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.*

Fan, C.; Liu, J.; Zhang, Y.; Wei, D.; Wong, E.; and Liu, S. 2024. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *ICLR*.

Foster, J.; Schoepf, S.; and Brintrup, A. 2024. Fast machine unlearning without retraining through selective synaptic dampening. In *AAAI*.

Goel, S.; Prabhu, A.; Sanyal, A.; Lim, S.-N.; Torr, P.; and Kumaraguru, P. 2022. Towards adversarial evaluations for inexact machine unlearning. *arXiv preprint arXiv:2201.06640*.

Golatkar, A.; Achille, A.; and Soatto, S. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *CVPR*.

Graves, L.; Nagisetty, V.; and Ganesh, V. 2021. Amnesiac machine learning. In *AAAI*.

Hayes, J.; Shumailov, I.; Triantafillou, E.; Khalifa, A.; and Papernot, N. 2024. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. *arXiv preprint arXiv:2403.01218*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

Hinton, G.; Vinyals, O.; and Dean, J. 2014. Distilling the knowledge in a neural network. *NIPS Workshop*.

Jeon, D.; Jeung, W.; Kim, T.; No, A.; and Choi, J. 2025. An information theoretic evaluation metric for strong unlearning. *arXiv preprint arXiv:2405.17878*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

Jia, J.; Liu, J.; Ram, P.; Yao, Y.; Liu, G.; Liu, Y.; Sharma, P.; and Liu, S. 2023. Model Sparsity Can Simplify Machine Unlearning. In *NeurIPS*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. In *NeurIPS*.

Kim, H.; Lee, S.; and Woo, S. S. 2024. Layer Attack Unlearning: Fast and Accurate Machine Unlearning via Layer Level Attack and Knowledge Distillation. In *AAAI*.

Kirichenko, P.; Izmailov, P.; and Wilson, A. G. 2023. Last layer re-training is sufficient for robustness to spurious correlations. *ICLR*.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images.

Kurmanji, M.; Triantafillou, P.; Hayes, J.; and Triantafillou, E. 2023. Towards unbounded machine unlearning. In *NeurIPS*.

Mu, S.; and Klabjan, D. 2024. Rewind-to-Delete: Certified Machine Unlearning for Nonconvex Functions. *arXiv preprint arXiv:2409.09778*.

Nasr, M.; Shokri, R.; and Houmansadr, A. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *IEEE S&P*.

Neel, S.; Roth, A.; and Sharifi-Malvajerdi, S. 2021. Descent-to-delete: Gradient-based methods for machine unlearning. In *ALT*.

Nguyen, T. T.; Huynh, T. T.; Nguyen, P. L.; Liew, A. W.-C.; Yin, H.; and Nguyen, Q. V. H. 2022. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Poole, B.; Ozair, S.; Van Den Oord, A.; Alemi, A.; and Tucker, G. 2019. On variational bounds of mutual information. In *ICML*.

Poppi, S.; Sarto, S.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2024. Multi-Class Unlearning for Image Classification via Weight Filtering. *IEEE Intelligent Systems*.

Qiao, X.; Zhang, M.; Tang, M.; and Wei, E. 2024. Efficient Online Unlearning via Hessian-Free Recollection of Individual Data Statistics. *arXiv preprint arXiv:2404.01712*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Sekhari, A.; Acharya, J.; Kamath, G.; and Suresh, A. T. 2021. Remember what you want to forget: Algorithms for machine unlearning. *NeurIPS*.

Shaik, T.; Tao, X.; Xie, H.; Li, L.; Zhu, X.; and Li, Q. 2023. Exploring the landscape of machine unlearning: A survey and taxonomy. *arXiv preprint arXiv:2305.06360*.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell System Technical Journal*.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *S&P*.

Sommer, D. M.; Song, L.; Wagh, S.; and Mittal, P. 2022. Towards probabilistic verification of machine unlearning. *PETS*.

Tarun, A. K.; Chundawat, V. S.; Mandal, M.; and Kankanhalli, M. 2023a. Deep regression unlearning. In *ICML*.

Tarun, A. K.; Chundawat, V. S.; Mandal, M.; and Kankanhalli, M. 2023b. Fast yet effective machine unlearning. *IEEE Trans. Neural Netw. Learn. Syst.*

Thudi, A.; Deza, G.; Chandrasekaran, V.; and Papernot, N. 2022. Unrolling sgd: Understanding factors influencing machine unlearning. In *EuroS&P*.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *JMLR*.

Voigt, P.; and Von dem Bussche, A. 2017. The eu general data protection regulation (gdpr). *Springer International Publishing*.

Wu, Y.; Dobriban, E.; and Davidson, S. 2020. Deltagrad: Rapid retraining of machine learning models. In *ICML*.

Xu, H.; Zhu, T.; Zhang, L.; Zhou, W.; and Yu, P. 2023. Machine Unlearning: A Survey. *ACM Computing Surveys*.

Yan, H.; Li, X.; Guo, Z.; Li, H.; Li, F.; and Lin, X. 2022. ARCANE: An Efficient Architecture for Exact Machine Unlearning. In *IJCAI*.

Zhang, B.; Chen, Z.; Shen, C.; and Li, J. 2024a. Verification of machine unlearning is fragile. *ICML*.

Zhang, B.; Dong, Y.; Wang, T.; and Li, J. 2024b. Towards certified unlearning for deep neural networks. *ICML*.